Power-Law Degree Distribution in the Connected Component of a Duplication Graph

Philippe Jacquet
INRIA, Saclay – Île-de-France, France
philippe.jacquet@inria.fr

Krzysztof Turowski
Theoretical Computer Science Department, Jagiellonian University, Krakow, Poland
krzysztof.szymon.turowski@gmail.com

Wojciech Szpankowski
Center for Science of Information, Department of Computer Science, Purdue University, West Lafayette, IN, USA
spa@cs.purdue.edu

Abstract

We study the partial duplication dynamic graph model, introduced by Bhan et al. in [3] in which a newly arrived node selects randomly an existing node and connects with probability p to its neighbors. Such a dynamic network is widely considered to be a good model for various biological networks such as protein-protein interaction networks. This model is discussed in numerous publications with only a few recent rigorous results, especially for the degree distribution. In particular, recently Jordan [9] proved that for $0 < p < \frac{1}{e}$ the degree distribution of the *connected component* is stationary with *approximately* a power law. In this paper we rigorously prove that the tail is indeed a true power law, that is, we show that the degree of a randomly selected node in the connected component decays like C/k^{β} where C an explicit constant and $\beta \neq 2$ is a non-trivial solution of $p^{\beta-2} + \beta - 3 = 0$. To establish this finding we apply analytic combinatorics tools, in particular Mellin transform and singularity analysis.

2012 ACM Subject Classification Mathematics of computing → Random graphs; Theory of computation → Random network models

Keywords and phrases random graphs, pure duplication model, degree distribution, tail exponent, analytic combinatorics

Digital Object Identifier 10.4230/LIPIcs.CVIT.2016.23

Funding This work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, in addition by NSF Grant CCF-1524312, and National Science Center, Poland, Grant 2018/31/B/ST6/01294.

42nd Conference on Very Important Topics (CVIT 2016).
Editors: John Q. Open and Joan R. Access; Article No. 23; pp. 23:1–23:15
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Recent years have seen a growing interest in dynamic graph models [10]. These models are often claimed to describe well various real-world structures, such as social networks, citation networks and various biological data. For instance, protein-protein and citation networks are widely viewed as driven by an internal evolution mechanism based on duplication and mutation. In this case, new nodes are added to the network as copies of existing nodes together with some random divergence. It has been claimed that graphs generated from these models exhibit many properties characteristic for real-world networks such as power-law degree distribution, the large clustering coefficient, and a large amount of symmetry [4]. However, some of these results turned out not to be correct; in particular, the power-law degree distribution was disproved in [7]. In this paper we focus on the tail distribution of the *connected component* of such networks and show rigorously the existence of a power law improving and making more precise recent result of Jordan [9].

To focus, we study here one of the more interesting models in this area known as the *partial (pure) duplication model*, in which a new node selects an existing node and connects to its neighbors with probability p. More precisely, the model is defined formally as follows: let $0 \leq p \leq 1$ be the only parameter of the model. In discrete steps we repeat the following procedure: first, we choose a single vertex u uniformly at random. Then, we add a new vertex v and for all vertices w such that uw is an edge (i.e., w is a neighbor of u) we flip a coin independently at random (heads with probability p, tails with $1-p$) and we add vw edge if and only if we got heads. The partial duplication model was defined by Bhan et al. in [3] and then was further studied in [1, 4, 7, 9, 8].

The case when $p = 1$, also called the *full duplication model*, was analyzed recently in the context of graph compression in [13]. In particular, it was formally proved that the expected logarithm of the number of automorphisms (symmetries) for such graphs on n vertices is asymptotically $\Theta(n \log n)$, which is a quite surprising result.

The *partial duplication* case $0 < p < 1$ was given much more attention, however, with very few rigorous results. It was first and foremost analyzed to find the stationary distribution of the degree, that is,

$$f_k = \lim_{n\to\infty} f_k(n) = \lim_{n\to\infty} \frac{F_k(n)}{n} = \lim_{n\to\infty} \Pr[\deg(U_n) = k],$$

where $F_n(k)$ is the average number of vertices of degree k in a graph generated by this model and U_n is a random variable denoting vertex chosen uniformly at random from a graph on n vertices generated from the partial duplication model. Hermann and Pfaffelhuber in [7] proved that there exist a limit with $f_0 = 1$ and $f_k = 0$ for all other k when $p \leq p^* = 0.58\dots$ (that is, p^* being the unique root of $pe^p = 1$). They have also shown that if $p > p^*$ there exists only a defective distribution of the degrees with $a_0 = c < 1$ for a certain constant c (depending on the initial graph) and $f_k = 0$ for all other k. For the average degree degree distribution see also [14].

This result, although it refuted the power law behavior of the whole graph claimed by [4, 2], still left the possibility that it might be the case for the large connected component of a graph generated by the partial duplication model. Note that by a simple inductive argument it is obvious that in such a graph there can only be one component containing non-isolated vertices, so there is no ambiguity. This was exactly the route pursued by Jordan in [9]. Using probabilistic tools such as the quasi-stationary distribution of a certain continuous time Markov chain embedding of the original discrete graph growth process, Jordan was able to prove that for $0 < p < \frac{1}{e}$ there is an *approximate power law* behavior in the pure duplication graphs. More precisely, let us define for a vertex (denoted by U_n) picked uniformly at random from a connected component of a graph on n vertices generated from the duplication model the following conditional probability

$$a_k(n) = \Pr[\deg(U_n) = k | \deg(U_n) \neq 0] = \frac{f_k(n)}{\sum_{i=1}^{\infty} f_i(n)} = \frac{f_k(n)}{1 - f_0(n)}. \tag{1}$$

Jordan proved that $a_k(n) \to a_k$ as $n \to \infty$ as long as the underlying process is positive recurrent which holds for for $p < \frac{1}{e}$ [9]. Moreover, Jordan showed that for $\beta(p) \neq 2$ being a solution of $p^{\beta-2} + \beta - 3 = 0$ the tail behavior of a_k is approximately a power law in the sense that it is lighter than any heavier tailed power law (with any index $\beta(p) + \varepsilon$, $\varepsilon > 0$) and heavier than any lighter tailed power law (with index $\beta(p) - \varepsilon$, $\varepsilon > 0$).

It is worth noting that it partially confirmed the non-rigorous result by Ispolatov et al. from [8], who claimed that the connected component exhibits a power-law distribution both for $0 < p < \frac{1}{e}$ (with index $\beta(p)$ as above), and for $\frac{1}{e} \leq p < \frac{1}{2}$ (with index 2). Furthermore, by the virtue of (1) we observe following [9, 7] that $f_0(n) = 1 - o(1)$ and $f_k(n) = o(1)$ for $k \geq 1$ which begs the question of the asymptotic behavior of $f_k(n)$ and $F_k(n)$ for large k and n. We can only say for certainty that $f_k(n)$ does *not* grow linearly with n as suggested in some papers (cf. [2]). We conjecture that $F_k(n) = O(n^{-\alpha} k^{-\beta})$ for some $0 < \alpha < 1$ and $\beta > 2$. We leave this problem for future research.

In this paper we finally establish the precise behavior of the tail of the degree distribution for pure duplication model for $0 < p < \frac{1}{e}$ completing the work of Jordan [9]. More precisely, we use tools of analytic combinatorics such as the Mellin transform and singularity analysis to prove in Theorem 2 that the tail of a node degree in the connect component of the partial duplication model decays as C/k^{β} where C an explicit constant and $\beta > 2$ is a non-trivial solution of $p^{\beta-2} + \beta - 3 = 0$.

The paper is organized as follows: in Section 2 we present a formal definition of the model, introduce the tracked vertex approach, and the quasi-stationary distribution as defined by Jordan in [9]. In Section 3 we state our results and using Mellin transform and singularity analysis we establish our main results. In concluding Section 4 we indicate a possible extension of our findings and pointing to some further work.

2 The model and Jordan's approach

We follow the standard graph-theoretical notation, e.g. from [5]. We consider only simple graphs, i.e. without loops or parallel edges.

Let us recall first the definition of the pure duplication model. Let $G_{n_0} = (V_{n_0}, E_{n_0})$ be an initial graph with a set of vertices V_{n_0} and a set of edges E_{n_0}, such that $|V_{n_0}| = n_0$. Throughout the paper, we assume that G_{n_0} is fixed and connected. For $n = n_0, n_0 + 1, \ldots$ we build $G_{n+1} = (V_{n+1}, E_{n+1})$ from $G_n = (V_n, E_n)$ in the following way:

1. pick a vertex $u \in V_n$ uniformly at random,
2. create a new node v_{n+1} and let $V_{n+1} = V_n \cup \{v_{n+1}\}$, $E_{n+1} = E_n$,
3. for every $w \in V_n$ such that $uw \in E_n$ add edge $v_{n+1}w$ to E_{n+1} independently at random with probability p.

We call the process $\mathcal{G} = (G_n)_{n=n_0}^{\infty}$ the *partial duplication graph.*

Jordan in [9] introduced the continuous-time embedding of this process, defined as following: we start at time 0 with a fixed connected graph $\Gamma_0 = G_0$ and let $(\Gamma_t)_{t \geq 0}$ be a continuous time Markov chain on graphs, where each vertex is duplicated independently at times following a Poisson process of rate 1, with the rules for duplication as in the pure duplication model.

Jordan also defined the so called *vertex tracking approach*: we pick a vertex from Γ_0 uniformly at random and then define the process $(V_t)_{t\geq 0}$ in the following way: at time t we jump to a vertex v if and only if the vertex V_{t-} was duplicated and its „child" is v. He proved that for any $k \geq 1$ and for $(U_t)_{t\geq 0}$ being defined as a uniform choice of vertices over Γ_t we have

$$\lim_{t\to\infty} \frac{\Pr[\deg(U_t) = k]}{\Pr[\deg(V_t) = k]} = 1.$$

Therefore, asymptotically the behavior of a tracked vertex approximates the behavior of a random vertex in Γ_t when $t \to \infty$, and therefore in G_n when $n \to \infty$.

The tracked vertex approach allowed Jordan to construct the generator Q of the continuous-time Markov chain $(\deg(V_t))_{t\geq 0}$, defined over the state space $\mathbb{N}_0$, with the following transitions

$$\begin{aligned} q_{j,k} &= \binom{j}{k} p^k (1-p)^{j-k} && \text{for } 0 \leq k \leq j-1, \\ q_{j,j} &= -jp - \left(1 - p^j\right), \\ q_{j,j+1} &= jp. \end{aligned}$$

Then we may proceed to the analysis of the quasi-stationary distribution $(a_k)_{k=1}^{\infty}$, i.e. the left eigenvector of a subset of Q, defined as $a_k = \lim_{n\to\infty} \Pr[\deg(U_n) = k | \deg(U_n) \neq 0]$. We relate this distribution to the eigenvalue $-\lambda$ (see [11] for details of this approach) being the solution of the equation $AQ = -\lambda Q$, where $A = (a_k)_{k=1}^{\infty}$. This leads us to the following equation:

$$\sum_{j=k}^{\infty} a_j \binom{j}{k} p^k (1-p)^{j-k} = -(k-1)pa_{k-1} - (\lambda - kp - 1)\, a_k \tag{2}$$

for $k = 1, 2, 3, \ldots$.

Using (2) and the generating function $A(z) = \sum_{k=0}^{\infty} a_k z^k$ Jordan found the following differential-functional equation

$$A(pz + 1 - p) = (1 - \lambda)A(z) + pz(1 - z)A'(z) + A(1 - p). \tag{3}$$

We notice that the above equation implies that $A(0) = 0$. Since it is a probability distribution, the function $A(z)$ exists for at least $|z| \leq 1$. By letting $z \to 1^-$ in (3) and assuming finite $A'(1)$ we get $A(1 - p) = \lambda A(1)$.

Furthermore with the identity

$$A'(z) = \frac{A(pz + 1 - p) - A(1)}{pz(1 - z)} - (1 - \lambda)\frac{A(z) - A(1)}{pz(1 - z)} \tag{4}$$

and letting $z \to 1^-$ Jordan found

$$A'(1) = -A'(1) + \frac{1 - \lambda}{p} A'(1),$$

namely, if $A'(1)$ is non-zero and finite, then $\lambda = 1 - 2p$. Finally, using the assumptions that the distribution $(a_k)_{k=0}^{\infty}$ is non-degenerate (i.e., $A(1) = 1$) and that the mean degree $A'(1)$ is finite, Jordan found that for $0 < p < \frac{1}{e}$ the quasi-stationary distribution a_k does not have q-th moment for $p^{q-2} + q - 3 < 0$.

In summary Jordan proved in [9] the following result.

▶ **Theorem 1** ([9, Theorem 2.1(3)]). *Assume* $0 < p < \frac{1}{e}$. *Let* $\beta(p)$ *be the solution of* $p^{\beta-2} + \beta - 3 = 0$. *Then the tail behaviour of* $(a_k)_{k=0}^{\infty}$ *has a power law of index* $\beta(p)$, *in the sense that as* $k \to \infty$,

$$\lim_{k\to\infty} \frac{a_k}{k^q} = 0 \quad \text{for } q < \beta(p),$$
$$\lim_{k\to\infty} \frac{a_k}{k^q} = \infty \quad \text{for } q > \beta(p).$$

We should note that, although it's missing from the statement of the theorem, $\beta(p)$ is supposed to be non-trivial solution of $p^{\beta-2} + \beta - 3 = 0$, i.e. other than $\beta = 2$. It may be checked that for the whole range $0 < p < \frac{1}{e}$ it is guaranteed to be unique.

In the next section we present our refinement of this theorem and provide precise asymptotics for $(a_k)_{k=0}^{\infty}$.

3 Main results

In this section we state and prove the main result of our paper that is a refinement of Theorem 1.

▶ **Theorem 2.** *If* $0 < p < \frac{1}{e}$, *then the stationary distribution* $(a_k)_{k=0}^{\infty}$ *of the pure duplication model has the following asymptotic tail behavior as* $k \to \infty$:

$$\frac{a_k}{k^{\beta(p)}} = \frac{1}{E(1) - E(\infty)} \cdot \frac{p^{-\frac{1}{2}(\beta(p)-\frac{3}{2})^2}\Gamma(\beta(p)-2)}{D(\beta(p)-2)(p^{-\beta(p)+2} + \ln(p))\Gamma(-\beta(p)+1)} \left(1 + O\left(\frac{1}{k}\right)\right) \tag{5}$$

where $\beta(p) > 2$ *is the non-trivial solution of* $p^{\beta-2} + \beta - 3 = 0$, $\Gamma(s)$ *is the Euler gamma function and*

$$D(s) = \prod_{i=0}^{\infty} \left(1 + p^{1+i-s}(s - i - 2)\right), \tag{6}$$
$$E(1) - E(\infty) = \frac{1}{2\pi i} \int_{\mathrm{Re}(s)=c} p^{-\frac{1}{2}(s-\frac{1}{2})^2} \frac{\Gamma(s)}{D(s)} \mathrm{d}s, \quad \text{for } c \in (0,1).$$

As we see from Figure 1, all coefficients in (3) are positive for $0 < p < \frac{1}{e}$.

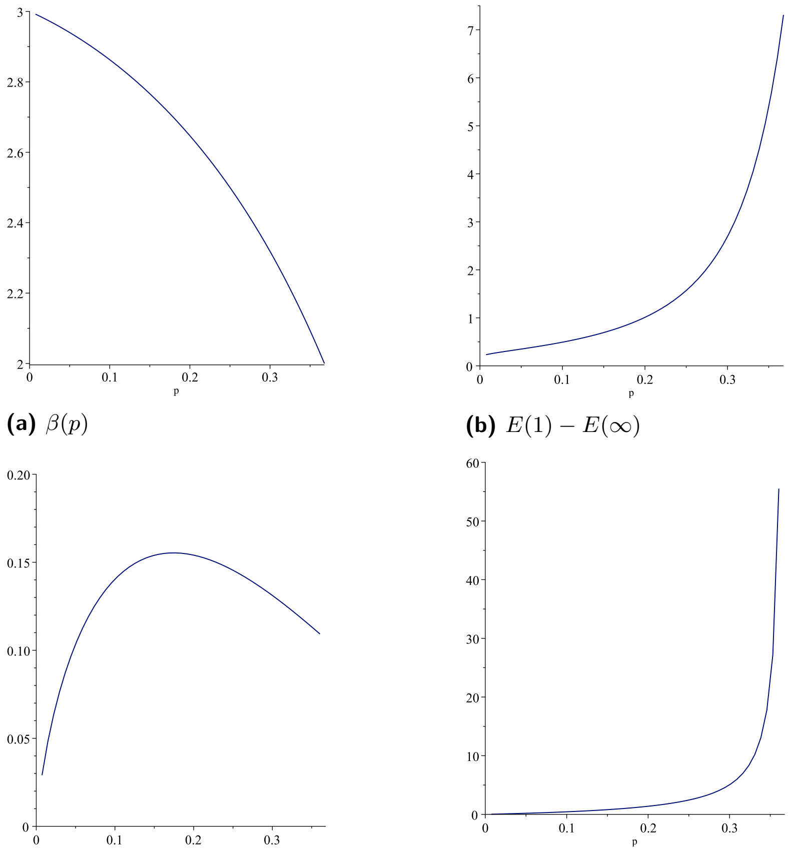
The rest of this section is devoted to the proof of our main result. We will accomplish it by a series of lemmas. The main idea is as follows: we take (3) and apply a series of substitutions to obtain a functional equation which is in suitable form for applying Mellin transform. Observe that we cannot apply directly Mellin transform to the functional equation (3) due to the term $A(pz + 1 - p)$.

We already know that if $A'(1) < \infty$ then $\lambda = 1 - 2p$. We first substitute $z = 1 - v$ and $B(v) = A(1 - v)$ in (3). Thus

$$A(1 - pv) = 2pA(1 - v) + pv(1 - v)A'(1 - v) + A(1 - p),$$
$$B(pv) = 2pB(v) - pv(1 - v)B'(v) + A(1 - p).$$

Observe now that the functional equation on $B(v)$ is suitable for the Mellin transform. However, to ease some computation we further let $w = \frac{1}{v}$ and $C(w) = B\left(\frac{1}{w}\right)$. Then

$$B\left(\frac{p}{w}\right) = 2pB\left(\frac{1}{w}\right) - \frac{p}{w}\left(1 - \frac{1}{w}\right) B'\left(\frac{1}{w}\right) + A(1 - p),$$



(a) $\beta(p)$

(b) $E(1) - E(\infty)$

(c) $D(\beta(p) - 2)$

(d) $\frac{p^{-\frac{1}{2}(\beta(p)-\frac{3}{2})^2}\Gamma(\beta(p)-2)}{(p^{-\beta(p)+2}+\ln(p))\Gamma(-\beta(p)+1)}$

Figure 1 Numerical values of different parts of (3) for $0 < p < \frac{1}{e}$.

$$C\left(\frac{w}{p}\right) = 2pC(w) + p(w-1)C'(w) + A(1-p). \tag{7}$$

Therefore, we are essentially looking at a solution of (7) with boundary conditions $C(1) = A(0) = 0$ and $\lim_{w\to\infty} C(w) = A(1)$ (which is equal to 1, as pointed out in [9]).

Our objective is to find an asymptotic expansion for $C(w)$ when $w \to \infty$. Notice that it is equivalent of finding the asymptotic expansion of $A(z)$ when $z \to 1$ by inferior values. For this purpose we will use the Mellin transform which is a powerful tool for extracting accurate asymptotic expansions [12]. Unfortunately we cannot directly apply the Mellin transform over function $C(w)$ since we do not know the behavior of $C(w)$ for $w \to 0$. To circumvent this problem we search for a similar function $E(w)$ defined by the following functional equation

$$E\left(\frac{w}{p}\right) = 2pE(w) + p(w-1)E'(w) + K \tag{8}$$

for some constant K for which the Mellin transform

$$E^*(s) = \int_0^\infty w^{s-1}E(w)\,\mathrm{d}w$$

exists in some fundamental strip.

To connect $E(w)$ with our function $C(w)$ we notice that we necessarily have $C(1) = 0$ which corresponds to the fact that $A(0) = 0$. Clearly, if we know that $E(w)$ is a solution of (8) with some finite values of $E(1)$ and denoting $E(\infty) = \lim_{w\to\infty} E(w)$, then we know also that

$$C(w) = A(1)\frac{E(w) - E(1)}{E(\infty) - E(1)} \tag{9}$$

is a solution of (7) with $C(1) = 0$ which also satisfies $\lim_{w\to\infty} C(w) = A(1) = 1$.

Let us now proceed though definition and lemmas. We first define

$$E^*(s) = p^{-\frac{1}{2}(s-\frac{1}{2})^2}\frac{\Gamma(s)}{D(s)} \tag{10}$$

for $D(s) = \prod_{i=0}^{\infty}\left(1 + p^{1+i-s}(s - i - 2)\right)$ defined already in (6).

Now we notice that $D(s) = 0$ only if $1 + p^{1+i-s}(s - i - 2) = 0$ for some $i \in \mathbb{N}$. This equation for $0 < p < \frac{1}{e}$ has only two solutions: $s = i + 1$ and $s = i + 1 + s^*$, where s^* is the non-trivial (i.e. other than $s = 0$) solution of $p^s + s - 1 = 0$.

Therefore, $E^*(s)$ has only isolated poles of three types:

- for $s = 0, -1, -2, \ldots$, introduced by $\Gamma(s)$,
- for $s = 1, 2, 3, \ldots$, introduced by $\frac{1}{D(s)}$,
- for $s = s^* + 1, s^* + 2, s^* + 3, \ldots$, introduced by $\frac{1}{D(s)}$.

Moreover, if we omit isolated poles, then $D(s)$ converges to a non-zero finite value when $\mathrm{Re}(s) < 0$ because p^{i-s} exponentially decays. We summarize it in the next lemma.

▶ **Lemma 3.** *For* $\mathrm{Re}(s) \in (-1, 0)$ *and* $0 < p < \frac{1}{e}$ *we have* $\frac{1}{|D(s)|}$ *absolutely convergent.*

Due to its technical intricacies, the proof of Lemma 3 was moved to the Appendix. In Figure 2 we present an example plot of $\frac{1}{|D(s)|}$.

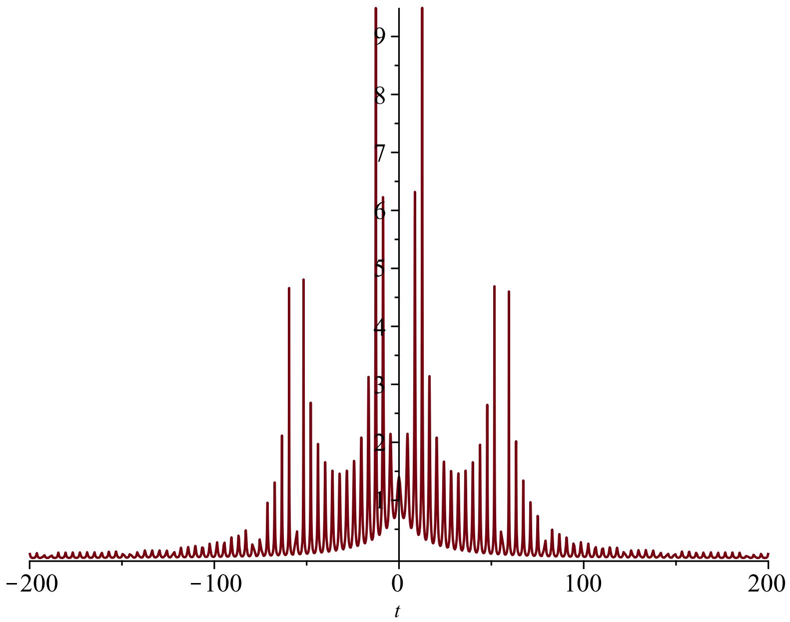


Figure 2 Numerical values of $\frac{1}{|D(c+it)|}$ for $p = 0.2$ and $c = -0.5$.

▶ **Lemma 4.** *It holds that*

$$E^*(s) = \frac{p(s-1)}{p^s + ps - 2p} E^*(s-1).$$

Proof. We have the identity

$$\frac{p^{\frac{1}{2}(s-\frac{1}{2})^2}}{\Gamma(s)} E^*(s) = \frac{p^{\frac{1}{2}(s-\frac{3}{2})^2}}{\Gamma(s-1)} E^*(s-1) \frac{1}{1+p^{1-s}(s-2)}$$

Thus

$$E^*(s) = \frac{p^{-\frac{1}{2}(s-\frac{1}{2})^2 + \frac{1}{2}(s-\frac{3}{2})^2}}{1+p^{1-s}(s-2)} \frac{\Gamma(s)}{\Gamma(s-1)} E^*(s-1) = \frac{p^{1-s}}{1+p^{1-s}(s-2)} (s-1) E^*(s-1)$$

since $\frac{\Gamma(s)}{\Gamma(s-1)} = s-1$. Multiplying by numerator and denominator by p^s completes the proof. ◀

We now state that for any given $c \in (-1, 0)$

$$E(w) = \frac{1}{2\pi i} \int_{\mathrm{Re}(s)=c} E^*(s) w^{-s} \, \mathrm{d}s = \frac{1}{2\pi i} \int_{\mathrm{Re}(s)=c} p^{-\frac{1}{2}(s-\frac{1}{2})^2} \frac{\Gamma(s)}{D(s)} w^{-s} \, \mathrm{d}s. \tag{11}$$

We notice that the integral converges for any complex value of w because from Lemma 3 it follows that $\frac{1}{|D(s)|}$ is bounded by a constant and $\Gamma(s) p^{-\frac{1}{2}(s-\frac{1}{2})^2}$ decays faster than any polynomial. Furthermore the value of $E(w)$ does not depends on the value of quantity c thanks to Cauchy theorem.

▶ **Lemma 5.** *The function $E(w)$ has function $E^*(s)$ as Mellin transform with its fundamental strip being $\{s : \mathrm{Re}(s) \in (-1, 0)\}$.*

Proof. We have

$$|E(w)| \leq \frac{|w|^{-c}}{2\pi} \int_{-\infty}^{+\infty} |E^*(c+it)| \exp(\arg(w) t) \, \mathrm{d}t.$$

Now, it is easy to spot that $E(c+it) = O\left(\exp\left(-\frac{t^2}{2}\right)\right)$ since $\ln(p) < -1$, thus the integral $\int_{-\infty}^{+\infty} |E^*(c+it)| \exp(\arg(w)t) \, \mathrm{d}t$ absolutely converges and it follows that $E(w) = O(w^{-c})$. Since it is true for any values of $c \in (-1, 0)$ when $w \to 0$ and $w \to \infty$, then the Mellin transforms of function $E(w)$ exists with the fundamental strip $\{s : \mathrm{Re}(s) \in (-1, 0)\}$.

Furthermore, its Mellin transform is $E^*(s)$ because (11) is exactly the inverse Mellin transform formula. ◀

▶ **Lemma 6.** *It holds that*

$$\mathrm{Res}\big[E^*(s-1) p(s-1) w^{-s}, s = 0\big] = -K.$$

Proof. The expression

$$R(w) = E\left(\frac{w}{p}\right) - 2pE(w) - p(w-1)E'(w)$$

can be also expressed via an integral as

$$R(w) = \frac{1}{2\pi i} \int_{\mathrm{Re}(s)=c} E^*(s) \left(p^s w^{-s} - 2pw^{-s} + spw^{-s} - spw^{-s-1}\right) \mathrm{d}s$$

which can be rewritten as following

$$R(w) = \frac{1}{2\pi i}\int_{\mathrm{Re}(s)=c} E^*(s)\,(p^s - 2p + ps)\,w^{-s}\,\mathrm{d}s$$

$$- \frac{1}{2\pi i}\int_{\mathrm{Re}(s)=c+1} E^*(s-1)p(s-1)w^{-s}\,\mathrm{d}s$$

$$= \frac{1}{2\pi i}\int_{\mathrm{Re}(s)=c} \left((p^s + ps - 2p)\,E^*(s) - p(s-1)E^*(s-1)\right)w^{-s}\,\mathrm{d}s$$

$$- \operatorname{Res}[p(s-1)E^*(s-1), s=0]$$

since

$$\int_{\mathrm{Re}(s)=c+1} p(s-1)E^*(s-1)w^{-s}\,\mathrm{d}s - \int_{\mathrm{Re}(s)=c} p(s-1)E^*(s-1)w^{-s}\,\mathrm{d}s$$

define a contour path which encircles a single pole at $s = 0$ in the counter-clockwise (i.e. positive) direction.

Furthermore from Lemma 5 we have

$$(p^s + ps - 2p)E^*(s) - p(s-1)E^*(s-1) = 0,$$

therefore the integral vanishes and we get $R(w) = -\operatorname{Res}[p(s-1)E^*(s-1), s=0] = K$. ◀

▶ **Lemma 7.** *We have*

$$K = \frac{p^{-\frac{1}{8}}(1-2p)}{D(0)}, \qquad E(\infty) = -\frac{p^{-\frac{1}{8}}}{D(0)}.$$

Furthermore,

$$E(\infty) - E(1) = -\frac{1}{2\pi i}\int_{\mathrm{Re}(s)=c} E^*(s)\,\mathrm{d}s, \quad \textit{for } c \in (0,1). \tag{12}$$

Proof. From Lemma 6 we infer that

$$K = -\operatorname{Res}[p(s-1)E^*(s-1), s=0] = \frac{p^{-\frac{1}{8}}}{D(-1)}.$$

Moreover, from the definition we know that $D(0) = (1-2p)D(-1)$, which establishes the first identity.

To find an expression for $E(\infty)$ is a little more delicate. Indeed we have from (11) the expression

$$E(w) = -\operatorname{Res}\left[E^*(s)w^{-s}, s=0\right] + \frac{1}{2\pi i}\int_{\mathrm{Re}(s)=c'} E^*(s)w^{-s}\,\mathrm{d}s$$

by assuming the contour path is moved right to origin for some $c' \in (0,1)$. It turns out that 0 is the single pole encountered in the move, as $D(s) \neq 0$ for all other s with $\mathrm{Re}(s) \in (0,1)$.

Furthermore, the integral on $\mathrm{Re}(s) = c'$ is in $O(w^{-c'})$, which allows to conclude that $E(w) = -\operatorname{Res}[E^*(s)w^{-s}, s=0] + O(w^{-c'})$ with $c' \in (0,1)$, thus

$$E(\infty) = \lim_{w\to\infty} E(w) = -\operatorname{Res}\left[E^*(s)w^{-s}, s=0\right] = -\operatorname{Res}[E(s), s=0] = -\frac{p^{-\frac{1}{8}}}{D(0)}.$$

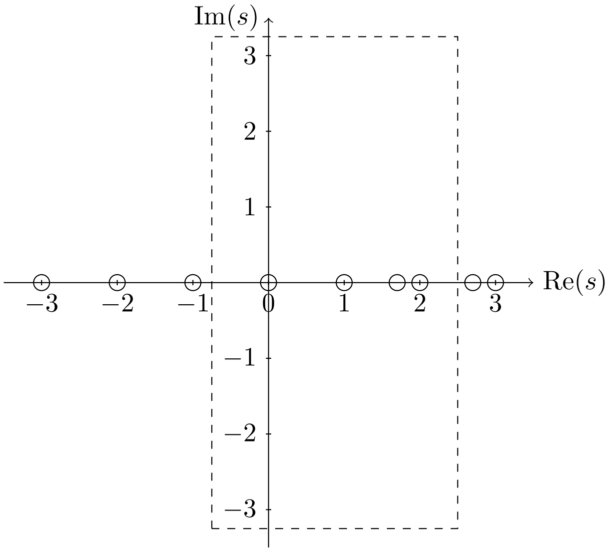


Figure 3 Example integration area for $E^*(s)$ and $E(w)$ with $s^* = 0.7$ and $M = 2.5$.

Finally,

$$E(\infty) - E(1) = -\operatorname{Res}[E(s), s = 0] - \frac{1}{2\pi i}\int_{\operatorname{Re}(s)=c} E^*(s)\,\mathrm{d}s = -\frac{1}{2\pi i}\int_{\operatorname{Re}(s)=c'} E^*(s)\,\mathrm{d}s$$

for, respectively, $c \in (-1, 0)$ and $c' \in (0, 1)$ since

$$\frac{1}{2\pi i}\int_{\operatorname{Re}(s)=c'} E^*(s)\,\mathrm{d}s - \frac{1}{2\pi i}\int_{\operatorname{Re}(s)=c} E^*(s)\,\mathrm{d}s = \operatorname{Res}[E(s), s = 0].$$

This completes the proof. ◀

We notice that $D(0) > 0$ since every element in the product is positive for $0 < p < \frac{1}{e}$. Therefore $K > 0$ and $E(\infty) < 0$.

Finally we proceed with the proof of our main theorem.

Proof of Theorem 2. We repeat the observation that $E^*(s)$ has poles for $s \in \{1, 2, \ldots\} \cup \{s^* + 1, s^* + 2, \ldots\} \cup \{0, -1, -2, \ldots\}$, for s^* – the non-zero solution of $p^s + s - 1 = 0$. Note that if $0 < p < \frac{1}{e}$, then $s^* > 0$.

Therefore, if we choose any $c \in (-1, 0)$ and draw a rectangle as presented in Figure 3, we are in position to write

$$\begin{aligned} C(w) &= \frac{1}{E(\infty) - E(1)}\frac{1}{2\pi i}\int_{\operatorname{Re}(s)=c} E^*(s)w^{-s}\,\mathrm{d}s - \frac{E(1)}{E(\infty) - E(1)} \\ &= -\frac{1}{E(\infty) - E(1)}\left(E(1) + \operatorname{Res}[E^*(s), s = 0] + \operatorname{Res}\left[E^*(s)w^{-s}, s = 1\right]\right) \\ &\quad - \frac{1}{E(\infty) - E(1)}\left(\operatorname{Res}\left[E^*(s)w^{-s}, s = 2\right] + \operatorname{Res}\left[E^*(s)w^{-s}, s = s^* + 1\right]\right) \\ &\quad + \frac{1}{E(\infty) - E(1)}\frac{1}{2\pi i}\int_{\operatorname{Re}(s)=M} E^*(s)w^{-s}\,\mathrm{d}s. \end{aligned} \tag{13}$$

for any number $M \in (2, 2 + s^*)$.

The quantity

$$\frac{1}{2\pi i}\int_{\mathrm{Re}(s)=M} E^*(s)w^{-s}\,\mathrm{d}s = O(w^{-M})$$

since $w^{-s}=w^{-M}w^{-\operatorname{Im}(s)}$ and the integral in $E^*(s)w^{-\operatorname{Im}(s)}$ absolutely converge. Again this holds by a similar argument that was used in Lemma 3: $p^{-\frac{1}{2}(s-\frac{1}{2})^2}$ decays exponentially faster than $\frac{\Gamma(s)}{D(s)}w^{\operatorname{Im}(s)}$ for complex s.

By virtue of the residue theorem we have

$$\begin{aligned} C(w) = &-\frac{1}{E(\infty)-E(1)}\left(E(1)+\mathrm{Res}[E^*(s), s=0]+\mathrm{Res}[E^*(s), s=1]w^{-1}\right) \\ &-\frac{1}{E(\infty)-E(1)}\left(\mathrm{Res}[E^*(s), s=2]w^{-2}+\mathrm{Res}[E^*(s), s=s^*+1]w^{-1-s^*}\right) \\ &+O(w^{-M}). \end{aligned} \tag{14}$$

This formula gives us an asymptotic expansion of $C(w)$ up to order w^{-M} where $M\in(2,2+s^*)$.

In fact, if we want more precise computations, we may write an expansion to any desired value M, just by including all the residues of the poles in k ($k\in\mathbb{N}$) and $k+s^*$ ($k\in\mathbb{N}_+$) which are smaller than M as we know that for $0<p<\frac{1}{e}$ all poles are single.

Next, we can compute the various residues, e.g.

$$\begin{aligned} \mathrm{Res}\big[E^*(s)w^{-s}, s=0\big] &= \left[p^{-\frac{1}{2}(s-\frac{1}{2})^2}\frac{w^{-s}}{D(s)}, s=0\right] = \frac{p^{-\frac{1}{8}}}{D(0)} = -E(\infty), \\ \mathrm{Res}\big[E^*(s)w^{-s}, s=1\big] &= \left[\frac{p^{-\frac{1}{2}(s-\frac{1}{2})^2}\Gamma(s)}{p^{1-s}-(s-2)p^{1-s}\ln(p)}\frac{w^{-s}}{D(s-1)}, s=1\right] \\ &= \frac{p^{-\frac{1}{8}}}{1+\ln(p)}\frac{w^{-1}}{D(0)}, \\ \mathrm{Res}\big[E^*(s)w^{-s}, s=s^*+1\big] &= \left[\frac{p^{-\frac{1}{2}(s-\frac{1}{2})^2}\Gamma(s)}{p^{1-s}-(s-2)p^{1-s}\ln(p)}\frac{w^{-s}}{D(s-1)}, s=s^*+1\right] \\ &= \frac{p^{-\frac{1}{2}(s^*+\frac{1}{2})^2}\Gamma(s^*)}{p^{-s^*}-(s^*-1)p^{-s^*}\ln(p)}\frac{w^{-s^*-1}}{D(s^*)} \\ &= \frac{p^{-\frac{1}{2}(s^*+\frac{1}{2})^2}\Gamma(s^*)}{p^{-s^*}+\ln(p)}\frac{w^{-s^*-1}}{D(s^*)}. \end{aligned}$$

Observe that in the formulas above both 1 and s^*+1 are not the zeros of $p^{1-s}-(s-2)p^{1-s}\ln(p)$, so all the presented expressions have finite value.

Now we are in position to use the classic Flajolet-Odlyzko transfer theorem [6] to (9) and (14) and obtain

$$\begin{aligned} A(z) = 1 &- \frac{1}{E(\infty)-E(1)}\frac{p^{-\frac{1}{8}}}{1+\ln(p)}\frac{1-z}{D(0)} \\ &- \frac{1}{E(\infty)-E(1)}\frac{p^{-\frac{1}{2}(s^*+\frac{1}{2})}\Gamma(s^*)}{p^{-s^*}+\ln(p)}\frac{(1-z)^{1+s^*}}{D(s^*)} \\ &- \frac{1}{E(\infty)-E(1)}\mathrm{Res}[E^*(s), s=2](1-z)^2 \\ &- \frac{1}{E(\infty)-E(1)}\mathrm{Res}[E^*(s), s=s^*+2](1-z)^{s^*+2}+o((1-z)^{2+s^*}). \end{aligned}$$

Finally, we use the fact that $(1-z)^\alpha$ for $\alpha \in \mathbb{N}$ is a polynomial and does not contribute to the asymptotics. And for $\alpha \in \mathbb{R}_+ \setminus \mathbb{N}$ [6] we have

$$[z^k](1-z)^\alpha = \frac{k^{-\alpha-1}}{\Gamma(-\alpha)}\left(1 + O\left(\frac{1}{k}\right)\right),$$
$$[z^k]o(1-z)^\alpha = o(k^{-\alpha-1}).$$

This lead us to the final result, which holds for large k:

$$\begin{aligned} a_k &= [z^k]A(z) \\ &= -\frac{1}{E(\infty)-E(1)}\frac{p^{-\frac{1}{2}(s^*+\frac{1}{2})^2}\Gamma(s^*)}{(p^{-s^*}+\ln(p))\Gamma(-s^*-1)}\frac{1}{D(s^*)}k^{-s^*-2}\left(1+O\left(\frac{1}{k}\right)\right). \end{aligned}$$

Note that since s^* is the non-trivial real solution of $p^s + s - 1 = 0$, we may equivalently write the exponent as $\beta(p) = s^* + 2$ – the the non-trivial (i.e. other than $\beta(p) = 2$) real solution of the equation $p^{\beta-2} + \beta - 3 = 0$.

Putting all the results together we obtain (5) of Theorem 2. Now it is sufficient to confirm that if $0 < p < \frac{1}{e}$, then the tail exponent $\beta(p) > 2$, which means that $A'(1)$ is indeed finite. This proves Theorem 2. ◀

4 Discussion

We observe that an eventual extension of our analysis for $\lambda = 1 - 2p$ up to $p < \frac{1}{2}$ would be almost identical, since the residues remain the same. The only difference would be that $s^* \in (-1, 0)$, therefore the fundamental strip would need to be taken more cautiously, with $c \in (-1, s^*)$. Therefore, our analysis might explain the non-rigorous claim in [8] that for $\frac{1}{e} < p < \frac{1}{2}$ the index of the power law is equal to 2.

However, for $\frac{1}{e} < p < \frac{1}{2}$ a formal application of residue theorem would result in $a_k \sim k^{-\beta}$ with $\beta < 2$ leading to $\sum_{k=0}^{\infty} k^q a_k = \infty$ for $q \geq 1$, which would violate the condition that $A'(1)$ is finite. We also note the fact that Jordan [9, Proposition 3.7] has shown that the dual Markov chain with respect to the eigenvalue $\lambda = 1 - 2p$ is transient for all $p > \frac{1}{e}$.

However, when $A'(1)$ is infinite we indicate there might be a possible way to extend our analysis. In particular, we present the following lemma.

▶ **Lemma 8.** *For the pure duplication model with $0 < p < \frac{1}{2}$ it holds that*

$$\lambda = 1 - p(1-\alpha) - p^{1-\alpha}$$

where α is such that $\lim_{z\to 1}(1-z)^\alpha A'(1)$ is finite and non-zero.

Proof. Assume that for $A'(1) = \sum_{k=1}^{\infty} k a_k$ (the expected degree of the quasi-stationary distribution) it holds that $\lim_{z\to 1}(1-z)^\alpha A'(z)$ is finite and non-zero for some $0 \leq \alpha < 1$. Then from the (3) we have

$$\begin{aligned} \lim_{z\to 1}(1-z)^\alpha A'(z) &= \lim_{z\to 1}\frac{A(pz+1-p)-(1-\lambda)A(z)-A(1-p)}{pz(1-z)^{1-\alpha}} \\ &= \lim_{z\to 1}\frac{p^{1-\alpha}(1-(pz+1-p))^\alpha A'(pz+1-p)-(1-\lambda)(1-z)^\alpha A'(z)}{p(1-z)+pz(\alpha-1)} \end{aligned}$$

which completes the proof. ◀

References

1. Gürkan Bebek, Petra Berenbrink, Colin Cooper, Tom Friedetzky, Joseph Nadeau, and Süleyman Cenk Sahinalp. The degree distribution of the generalized duplication model. *Theoretical Computer Science*, 369(1-3):239–249, 2006.
2. Gürkan Bebek, Petra Berenbrink, Colin Cooper, Tom Friedetzky, Joseph Nadeau, and Süleyman Cenk Sahinalp. The degree distribution of the generalized duplication model. *Theoretical Computer Science*, 369(1-3):239–249, 2006.
3. Ashish Bhan, David Galas, and T. Gregory Dewey. A duplication growth model of gene expression networks. *Bioinformatics*, 18(11):1486–1493, 2002.
4. Fan Chung, Linyuan Lu, T. Gregory Dewey, and David Galas. Duplication models for biological networks. *Journal of Computational Biology*, 10(5):677–687, 2003.
5. Reinhard Diestel. *Graph Theory*. Springer, 2005.
6. Philippe Flajolet and Andrew Odlyzko. Singularity analysis of generating functions. *SIAM Journal on Discrete Mathematics*, 3(2):216–240, 1990.
7. Felix Hermann and Peter Pfaffelhuber. Large-scale behavior of the partial duplication random graph. *ALEA: Latin American Journal of Probability and Mathematical Statistics*, 13:687–710, 2016.
8. I. Ispolatov, P. L. Krapivsky, and A. Yuryev. Duplication-divergence model of protein interaction network. *Phys. Rev. E*, 71:061911, 2005. doi:10.1103/PhysRevE.71.061911.
9. Jonathan Jordan. The connected component of the partial duplication graph. *ALEA: Latin American Journal of Probability and Mathematical Statistics*, 15:1431–1445, 2018.
10. Mark Newman. *Networks: An Introduction*. Oxford University Press, 2010.
11. PK Pollett. Reversibility, invariance and μ-invariance. *Advances in applied probability*, 20(3):600–621, 1988.
12. Wojciech Szpankowski. *Average Case Analysis of Algorithms on Sequences*. John Wiley & Sons, New York, 2001.
13. Krzysztof Turowski, Abram Magner, and Wojciech Szpankowski. Compression of Dynamic Graphs Generated by a Duplication Model. In *56th Annual Allerton Conference on Communication, Control, and Computing*, pages 1089–1096, 2018.
14. Krzysztof Turowski and Wojciech Szpankowski. Towards degree distribution of duplication graph models, 2019.

A Proof of Lemma 3

We now proceed to the proof of Lemma 3. First, we introduce $f(s) = p^s + ps - 2p$, so that

$$D(s) = \prod_{i=0}^{\infty} f(s-i)p^{-(s-i)}.$$

Observe that $f(s)$ has only two roots, given by Lambert function W, which is the inverse of function xe^x: $W^{-1}(x) = xe^x$. There are only two roots for real numbers which corresponds to two branches W_0 and W_{-1} of the function W. Therefore, if we pick any $c < 0$, then it is smaller than the roots of $f(s)$ and the distance between c and any root is at least 1.

▶ **Lemma 9.** *For all $\varepsilon > 0$ and $c < 0$ we have $\min_{\mathrm{Re}(s)=c} |f(s)| \geq \Theta(p^{(1-\varepsilon)(c-1)}) > 0$.*

Proof. We have $f'(s) = p^s \ln(p) + p$ and $f''(s) = p^s \ln^2(p)$.

Let us consider a complex disk of radius $R = p^{-\varepsilon(c-1)}$ $(R < 1)$ centered on s. For $\theta \in (0, 2\pi)$ by virtue of Taylor-Young theorem we have:

$$f(s + Re^{i\theta}) = f(s) + f'(s)e^{i\theta}R + \int_0^R f''(s + \rho e^{i\theta})e^{2i\theta}\rho \,\mathrm{d}\rho.$$

We observe that

$$\begin{aligned}
\left| \int_0^R f''(s + \rho e^{i\theta})e^{2i\theta}\rho \,\mathrm{d}\rho \right| &= \left| p^s \ln^2(p)e^{2i\theta} \int_0^R p^{\rho \exp(i\theta)}\rho \,\mathrm{d}\rho \right| \\
&= \left| p^s \left(e^{R\exp(i\theta)\ln(p)} \left[R\exp(i\theta)\ln(p) - 1 \right] + 1 \right) \right| \\
&= O\left(|p^s R^2 e^{2i\theta}| \right) = O\left(p^c R^2 \right),
\end{aligned}$$

where the last line follows from the fact that asymptotically $e^x(x-1)+1 = O(x^2)$ for $x \to 0$.

When θ varies the quantity $f'(s)e^{i\theta}R$ describes a circle of radius $|f'(s)|R = -p^c \ln(p)R$ around $f(s)$. The error term bound implies that each point of $f(s + Re^{i\theta})$ is at distance $O(p^c R^2)$ of this circle. Thus the image by f of the disk with center s and radius R contains the disk of center $f(s)$ and radius

$$\begin{aligned}
R|f'(s)| - O(R^2 p^c) &= -p^{-\varepsilon(c-1)}p^c \ln(p) - O\left(p^{-2\varepsilon(c-1)}p^c\right) \\
&= p^{(1-\varepsilon)(c-1)} \left(-p\ln(p) - O(p^{-\varepsilon(c-1)}) \right) = \Theta(p^{(1-\varepsilon)(c-1)}).
\end{aligned}$$

We know that 0 cannot be in this disk, otherwise the function $f(s)$ would have other roots than the expected ones, thus we necessarily have $|f(s)| \geq \Theta(p^{(1-\varepsilon)(c-1)})$. ◀

Let now $g(s) = p^{-s}f(s)$ so that

$$D(s) = \prod_{i=0}^{\infty} g(s-i).$$

▶ **Lemma 10.** *For t real and $c < 0$, we have the following inequality*

$$|g(c+it)| \geq |1 - p^{1-c}(2-c) - p^{1-c}|t||.$$

Proof. We have

$$\begin{aligned}
|g(c+it)| &= |p^{-c}f(c+it)| = |p^{it} + p^{1-c}(c-2) + p^{1-c}it| \\
&\geq ||p^{it}| - |p^{1-c}(c-2)| - |p^{1-c}it||.
\end{aligned}$$

But now we observe that $|p^{it}| = 1$, which completes the proof. ◀

▶ **Lemma 11.** *For $c \in (-1,0)$ and for all real number t outside any neighborhood of 0, for all $\varepsilon > 0$ we have* $\frac{1}{D(c+it)} = O(\exp\left(-(\log_p^2|t|/2 + O(\log|t|)\right)$.

Proof. From Lemmas 9 and 10 we have, neglecting a term $p^{1-c+k}(2-c+k)$ which exponentially decays:

$$|D(c+it)| \geq \prod_{k\geq 0} \max\{Bp^{-\varepsilon(1-c)}, \left|1-|t|/p^{k+1-c}\right|\}$$

We denote $k(t) = -\lceil c + \log_p |t| \rceil$. To simplify we assume that $t = p^{c-k(t)}$, ie. $c + \log_p |t|$ is integer, we have

$$|D(c+it)| \geq \left(\prod_{k<k(t)} (|t|/t(c-k)-1)\right) B'|t|^{-\epsilon}\left(\prod_{k>k(t)}(1-|t|/t(c-k))\right)$$

$$\geq \left(\prod_{k<k(t)} (p^{-k}-1)\right) B'|t|^{-\epsilon}\prod_{k>0}(1-p^k) \tag{15}$$

Since $p^{-k}-1 = p^{-k}(1-p^k)$ we get $\prod_{k<k(t)}(p^{-k}-1) \geq p^{k(t)(k(t)-1)/2}\prod_{k>0}(1-p^k)$ and since $p^{-k(t)} = |t|p^{-c}$:

$$|D(c+it)| \geq p^{k(t)(k(t)-1)/2}B'|t|^{-\epsilon}\prod_{k>0}(1-p^k)^2 = B''\frac{|t|^{-\epsilon}}{(|t|p^{-c})^{(k(t)-1)/2}}.$$

We conclude, since $k(t) = c - \log_p |t|$. ◀

Notice that $D(c+it)$ tends to infinity when $|t| \to \infty$. To conclude the proof of Lemma 3 it is sufficient to observe that the function $1/D(s)$ for s is any compact set containing a neighborhood of $\mathrm{Re}(s)$ and away from the roots of $f(s)$ is naturally bounded by dominated convergence of the product.